(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
2 August 2001 (02.08.2001)

PCT

(10) International Publication Number
WO 01/55951 A2

(51) International Patent Classification[7]: G06F 19/00

(21) International Application Number: PCT/US01/02294

(22) International Filing Date: 24 January 2001 (24.01.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

60/177,964	25 January 2000 (25.01.2000)	US
60/201,105	2 May 2000 (02.05.2000)	US
09/768,686	24 January 2001 (24.01.2001)	US

(71) Applicant *(for all designated States except US)*: **CELLOMICS, INC.** [US/US]; 635 William Pitt Way, Pittsburgh, PA 15238 (US).

(72) Inventor; and

(75) Inventor/Applicant *(for US only)*: **BUSA, William, Brian** [US/US]; 201 Johns School Road, Renfrew, PA 16053 (US).

(74) Agent: **LESAVICH, Stephen**; McDonnell Boehnen Hulbert & Berghoff, Suite 3200, 300 South Wacker Drive, Chicago, IL 60606 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 01/55951 A2

(54) Title: METHOD AND SYSTEM FOR AUTOMATED INFERENCE CREATION OF PHYSICO-CHEMICAL INTERACTION KNOWLEDGE FROM DATABASES OF CO-OCCURRENCE DATA

(57) Abstract: Methods and system for automated inference of physico-chemical interaction knowledge from databases of term co-occurrence data. The co-occurrence data includes co-occurrences between chemical or biological molecules or co-occurrences between chemical or biological molecules and biological processes. Likelihood statistics are determined and applied to decide if co-occurrence data reflecting physico-chemical interactions is non-trivial. A next node or an unknown target representing chemical or biological molecules in a biological pathway is selected based on co-occurrence values. The method and system may be used to further facilitate a user's understanding of biological functions, such as cell functions, to design experiments more intelligently and to analyze experimental results more thoroughly. Specifically, the present invention may help drug discovery scientists select better targets for pharmaceutical intervention in the hope of curing diseases. The method and system may also help facilitate the abstraction of knowledge from information for biological experimental data and provide new bioinformatic techniques.

DERWENT-ACC-NO: 2001-476263

DERWENT-WEEK: 200279

TITLE: Strength measurement of co-occurrence data for automated interference of physico-chemical interaction knowledge, involves determining if co-occurrence between at least two chemical or biological molecule names is non-trivial

INVENTOR: BUSA, W B

PATENT-ASSIGNEE: CELLOMICS INC[CELLN] , BUSA W B[BUSAI]

PRIORITY-DATA: 2001US-0768686 (January 24, 2001) , 2000US-177964P (January 25, 2000) , 2000US-201105P (May 2, 2000) , 2001US-0769169 (January 24, 2001)

PATENT-FAMILY:

PUB-NO	PUB-DATE	LANGUAGE	PAGES	MAIN-IPC
EP 1252598 A2	October 30, 2002	E	000	G06F 019/00
WO 200155951 A2	August 2, 2001	E	063	G06F 019/00
AU 200129744 A	August 7, 2001	N/A	000	G06F 019/00
AU 200132928 A	August 7, 2001	N/A	000	G06F 019/00
US 20020002559 A1	January 3, 2002	N/A	000	G06F 017/30
US 20020004792 A1	January 10, 2002	N/A	000	G06N 005/02

DESIGNATED-STATES: AL AT BE CH CY DE DK ES FI FR GB GR IE IT LI LT LU LV MC MK NL PT RO SE SI TR AE AG AL AM AT AU AZ BA BB BG BR BY BZ CA CH CN CR CU CZ DE DK DM DZ EE ES FI GB GD GE GH GM HR HU ID IL IN IS JP KE KG KP KR KZ LC LK LR LS LT LU LV MA MD MG MK MN MW MX MZ NO NZ PL PT RO RU SD SE SG SI SK SL TJ TM TR TT TZ UA UG US UZ VN YU ZA ZW AT BE CH CY DE DK EA ES FI FR GB GH GM GR IE IT KE LS LU MC MW MZ NL OA PT SD SE SL SZ TR TZ UG ZW

APPLICATION-DATA:

PUB-NO	APPL-DESCRIPTOR	APPL-NO	APPL-DATE
EP 1252598A2	N/A	2001EP-0946969	January 24, 2001
EP 1252598A2	N/A	2001WO-US02294	January 24, 2001
EP 1252598A2	Based on	WO 200155951	N/A
WO 200155951A2	N/A	2001WO-US02294	January 24, 2001
AU 200129744A	N/A	2001AU-0029744	January 24, 2001
AU 200129744A	Based on	WO 200155951	N/A
AU 200132928A	N/A	2001AU-0032928	January 24, 2001
AU 200132928A	Based on	WO 200155950	N/A
US20020002559A1	Provisional	2000US-177964P	January 25, 2000
US20020002559A1	N/A	2001US-0769169	January 24, 2001
US20020004792A1	Provisional	2000US-177964P	January 25, 2000
US20020004792A1	Provisional	2000US-201105P	May 2, 2000

US20020004792A1 N/A 2001US-0768686 January 24, 2001

INT-CL (IPC): G06F017/30, G06F019/00 , G06N005/02

RELATED-ACC-NO: 2001-496878

ABSTRACTED-PUB-NO: US20020002559A

BASIC-ABSTRACT:

NOVELTY - A strength of co-occurrence data is measured by extracting at least two chemical or biological molecule names from database record; and determining likelihood statistic for co-occurrence reflecting physico-chemical interactions between the two molecule names, and applying it to the co-occurrence to determine if co-occurrence between the molecule names is non-trivial.

DETAILED DESCRIPTION - Strength measurement of co-occurrence data involves extracting at least two chemical or biological molecule names from database record from an interference database; determining likelihood statistic for co-occurrence reflecting physico-chemical interactions between the two molecule names (A and B); and applying the likelihood statistic to the co-occurrence to determine if the co-occurrence between molecule A and molecule B is non-trivial. The interference database includes those records created from an indexed literature database. The two molecule names co-occur in at least one record in an indexed scientific literature database.

An INDEPENDENT CLAIM is also included for:

(1) a method of contextual querying of co-occurrence data comprising selecting a target node from a first list of nodes connected by arcs in a connection network; creating a second list of nodes by considering other nodes that are neighbors of the target node and other nodes in prior to the target node in the connection network; selecting a next node from the second list of nodes using the co-occurrence values, in which the next node is next after the target node in the pre-determined order for the connection network based on the co-occurrence values;

(2) method of query polling of co-occurrence data comprising selecting a position in connection network for an unknown target node from a first list of nodes; determining a second list of nodes prior to the position of unknown target node in the connection network; determining a third list of nodes subsequent to the position of unknown target node in the connection network; determining a fourth list of nodes included in both the second and the third lists of nodes; and determining an identity for the unknown target node by selecting a node from the fourth list of nodes using likelihood statistic; and

(3) a method for creating automated biological interferences comprising constructing a connection network using at least one database record from an interference database; applying likelihood statistics analysis methods to the connection network; generating automatically at least one biological

interferences relationships between chemical or biological molecules or biological processes using the results from the likelihood statistic analysis methods.

USE - The method is for automated interference of physico-chemical interaction knowledge from databases of term co-occurrence data. It can also be used to facilitate a user's understanding of biological functions, e.g. cell functions, to design experiments, and to analyze experiment results.

ADVANTAGE - The method helps drug discovery scientists select better targets for pharmaceutical intervention of curing diseases. It may also help facilitate the abstraction of knowledge from information for biological experimental data and provides new bioinformatic techniques.

ABSTRACTED-PUB-NO: US20020004792A

EQUIVALENT-ABSTRACTS:

NOVELTY - A strength of co-occurrence data is measured by extracting at least two chemical or biological molecule names from database record; and determining likelihood statistic for co-occurrence reflecting physico-chemical interactions between the two molecule names, and applying it to the co-occurrence to determine if co-occurrence between the molecule names is non-trivial.

DETAILED DESCRIPTION - Strength measurement of co-occurrence data involves extracting at least two chemical or biological molecule names from database record from an interference database; determining likelihood statistic for co-occurrence reflecting physico-chemical interactions between the two molecule names (A and B); and applying the likelihood statistic to the co-occurrence to determine if the co-occurrence between molecule A and molecule B is non-trivial. The interference database includes those records created from an indexed literature database. The two molecule names co-occur in at least one record in an indexed scientific literature database.

An INDEPENDENT CLAIM is also included for:

(1) a method of contextual querying of co-occurrence data comprising selecting a target node from a first list of nodes connected by arcs in a connection network; creating a second list of nodes by considering other nodes that are neighbors of the target node and other nodes in prior to the target node in the connection network; selecting a next node from the second list of nodes using the co-occurrence values, in which the next node is next after the target node in the pre-determined order for the connection network based on the co-occurrence values;

(2) method of query polling of co-occurrence data comprising selecting a position in connection network for an unknown target node from a first list of nodes; determining a second list of nodes prior to the position of unknown target node in the connection network; determining a third list of nodes subsequent to the position of unknown target node in the connection network; determining a fourth list of nodes included in both the second and the third

lists of nodes; and determining an identity for the unknown target node by selecting a node from the fourth list of nodes using likelihood statistic; and

(3) a method for creating automated biological interferences comprising constructing a connection network using at least one database record from an interference database; applying likelihood statistics analysis methods to the connection network; generating automatically at least one biological interferences relationships between chemical or biological molecules or biological processes using the results from the likelihood statistic analysis methods.

USE - The method is for automated interference of physico-chemical interaction knowledge from databases of term co-occurrence data. It can also be used to facilitate a user's understanding of biological functions, e.g. cell functions, to design experiments, and to analyze experiment results.

ADVANTAGE - The method helps drug discovery scientists select better targets for pharmaceutical intervention of curing diseases. It may also help facilitate the abstraction of knowledge from information for biological experimental data and provides new bioinformatic techniques.

NOVELTY - A strength of co-occurrence data is measured by extracting at least two chemical or biological molecule names from database record; and determining likelihood statistic for co-occurrence reflecting physico-chemical interactions between the two molecule names, and applying it to the co-occurrence to determine if co-occurrence between the molecule names is non-trivial.

DETAILED DESCRIPTION - Strength measurement of co-occurrence data involves extracting at least two chemical or biological molecule names from database record from an interference database; determining likelihood statistic for co-occurrence reflecting physico-chemical interactions between the two molecule names (A and B); and applying the likelihood statistic to the co-occurrence to determine if the co-occurrence between molecule A and molecule B is non-trivial. The interference database includes those records created from an indexed literature database. The two molecule names co-occur in at least one record in an indexed scientific literature database.

An INDEPENDENT CLAIM is also included for:

(1) a method of contextual querying of co-occurrence data comprising selecting a target node from a first list of nodes connected by arcs in a connection network; creating a second list of nodes by considering other nodes that are neighbors of the target node and other nodes in prior to the target node in the connection network; selecting a next node from the second list of nodes using the co-occurrence values, in which the next node is next after the target node in the pre-determined order for the connection network based on the co-occurrence values;

(2) method of query polling of co-occurrence data comprising selecting a position in connection network for an unknown target node from a first list of nodes; determining a second list of nodes prior to the position of unknown

target node in the connection network; determining a third list of nodes subsequent to the position of unknown target node in the connection network; determining a fourth list of nodes included in both the second and the third lists of nodes; and determining an identity for the unknown target node by selecting a node from the fourth list of nodes using likelihood statistic; and

(3) a method for creating automated biological interferences comprising constructing a connection network using at least one database record from an interference database; applying likelihood statistics analysis methods to the connection network; generating automatically at least one biological interferences relationships between chemical or biological molecules or biological processes using the results from the likelihood statistic analysis methods.

USE - The method is for automated interference of physico-chemical interaction knowledge from databases of term co-occurrence data. It can also be used to facilitate a user's understanding of biological functions, e.g. cell functions, to design experiments, and to analyze experiment results.

ADVANTAGE - The method helps drug discovery scientists select better targets for pharmaceutical intervention of curing diseases. It may also help facilitate the abstraction of knowledge from information for biological experimental data and provides new bioinformatic techniques.

WO 200155951A

CHOSEN-DRAWING: Dwg.0/9

TITLE-TERMS: STRENGTH MEASURE CO OCCUR DATA AUTOMATIC INTERFERENCE PHYSICO
CHEMICAL INTERACT DETERMINE CO OCCUR TWO CHEMICAL BIOLOGICAL MOLECULAR NAME NON

DERWENT-CLASS: B04 D16 T01

CPI-CODES: B11-C08; B12-K04; D05-H09;

EPI-CODES: T01-J;

CHEMICAL-CODES:
Chemical Indexing M6 *01*
Fragmentation Code
M905 P831 Q010 Q233 R501 R511 R515 R528

SECONDARY-ACC-NO:
CPI Secondary Accession Numbers: C2001-142902
Non-CPI Secondary Accession Numbers: N2001-352481

(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) Veröffentlichungsnummer:
(11) Publication number: **EP 1 252 598 A0**
(11) Numéro de publication:

Internationale Anmeldung veröffentlicht durch die Weltorganisation für geistiges Eigentum unter der Nummer:

WO 01/055951 (art. 158 des EPÜ).

International application published by the World Intellectual Property Organisation under number:

WO 01/055951 (art. 158 of the EPC).

Demande internationale publiée par l'Organisation Mondiale de la Propriété sous le numéro:

WO 01/055951 (art. 158 de la CBE).

L Number	Hits	Search Text	DB	Time stamp
1	32	inference adj2 database	USPAT; US-PGPUB	2003/06/04 12:12
2	10	inference adj2 database	EPO; JPO; DERWENT; IBM_TDB	2003/06/04 12:12
3	94	database and literature and chemical and biological and structured and filter$3 and record	USPAT; US-PGPUB	2003/06/04 12:25
4	92	(database and literature and chemical and biological and structured and filter$3 and record) not (inference adj2 database)	USPAT; US-PGPUB	2003/06/04 12:16
5	0	database and literature and chemical and biological and structured and filter$3 and record	EPO; JPO; DERWENT; IBM_TDB	2003/06/04 12:26
6	0	database and literature and chemical and biological and filter$3 and record	EPO; JPO; DERWENT; IBM_TDB	2003/06/04 12:26
7	1	database and literature and chemical and biological and record	EPO; JPO; DERWENT; IBM_TDB	2003/06/04 12:28
8	1	2001-476263.NRAN.	DERWENT	2003/06/04 12:27
9	4	database and literature and chemical and biological	EPO; JPO; DERWENT; IBM_TDB	2003/06/04 12:28

```
Enter NEWS followed by the item number or name to see news on that
specific topic.

  All use of STN is subject to the provisions of the STN Customer
  agreement.  Please note that this agreement limits use to scientific
  research.  Use for software development or design or implementation
  of commercial gateways or other similar uses is prohibited and may
  result in loss of user privileges and other penalties.

  * * * * * * * * * * * * * * STN  Columbus  * * * * * * * * * * * * * * * *

FILE 'HOME' ENTERED AT 12:37:37 ON 04 JUN 2003

=> index bioscience
FILE 'DRUGMONOG' ACCESS NOT AUTHORIZED
COST IN U.S. DOLLARS                                    SINCE FILE      TOTAL
                                                           ENTRY    SESSION
FULL ESTIMATED COST                                         0.21       0.21

INDEX 'ADISCTI, ADISINSIGHT, ADISNEWS, AGRICOLA, ANABSTR, AQUASCI, BIOBUSINESS,
       BIOCOMMERCE, BIOSIS, BIOTECHABS, BIOTECHDS, BIOTECHNO, CABA, CANCERLIT,
       CAPLUS, CEABA-VTB, CEN, CIN, CONFSCI, CROPB, CROPU, DDFB, DDFU, DGENE,
       DRUGB, DRUGLAUNCH, DRUGMONOG2, ...' ENTERED AT 12:37:51 ON 04 JUN 2003

67 FILES IN THE FILE LIST IN STNINDEX

Enter SET DETAIL ON to see search term postings or to view
search error messages that display as 0* with SET DETAIL OFF.

=> s inference and database and literature and (chemical or biological)
          2   FILE BIOSIS
          1   FILE BIOTECHNO
          1   FILE CANCERLIT
          7   FILE CAPLUS
          2   FILE CEN
  18 FILES SEARCHED...
  26 FILES SEARCHED...
          1   FILE EMBASE
          2   FILE FEDRIP
          2   FILE IFIPAT
          1   FILE LIFESCI
          3   FILE MEDLINE
  47 FILES SEARCHED...
          2   FILE PASCAL
          4   FILE PROMT
          4   FILE SCISEARCH
        251   FILE USPATFULL
          2   FILE USPAT2
  66 FILES SEARCHED...

  15 FILES HAVE ONE OR MORE ANSWERS,   67 FILES SEARCHED IN STNINDEX

L1   QUE INFERENCE AND DATABASE AND LITERATURE AND (CHEMICAL OR BIOLOGICAL)

=> d rank
F1         251   USPATFULL
F2           7   CAPLUS
F3           4   PROMT
F4           4   SCISEARCH
F5           3   MEDLINE
F6           2   BIOSIS
```

```
F7            2   CEN
F8            2   FEDRIP
F9            2   IFIPAT
F10           2   PASCAL
F11           2   USPAT2
F12           1   BIOTECHNO
F13           1   CANCERLIT
F14           1   EMBASE
F15           1   LIFESCI


=> file f3-f15; s l1; dup rem l2; focus l3
COST IN U.S. DOLLARS                          SINCE FILE      TOTAL
                                                 ENTRY    SESSION
FULL ESTIMATED COST                               1.65       1.86

FILE 'PROMT' ENTERED AT 12:39:46 ON 04 JUN 2003
COPYRIGHT (C) 2003 Gale Group. All rights reserved.

FILE 'SCISEARCH' ENTERED AT 12:39:46 ON 04 JUN 2003
COPYRIGHT 2003 THOMSON ISI

FILE 'MEDLINE' ENTERED AT 12:39:46 ON 04 JUN 2003

FILE 'BIOSIS' ENTERED AT 12:39:46 ON 04 JUN 2003
COPYRIGHT (C) 2003 BIOLOGICAL ABSTRACTS INC.(R)

FILE 'CEN' ENTERED AT 12:39:46 ON 04 JUN 2003
COPYRIGHT (C) 2003 American Chemical Society (ACS)

FILE 'FEDRIP' ENTERED AT 12:39:46 ON 04 JUN 2003

FILE 'IFIPAT' ENTERED AT 12:39:46 ON 04 JUN 2003
COPYRIGHT (C) 2003 IFI CLAIMS(R) Patent Services (IFI)

FILE 'PASCAL' ENTERED AT 12:39:46 ON 04 JUN 2003
Any reproduction or dissemination in part or in full,
by means of any process and on any support whatsoever
is prohibited without the prior written agreement of INIST-CNRS.
COPYRIGHT (C) 2003 INIST-CNRS. All rights reserved.

FILE 'USPAT2' ENTERED AT 12:39:46 ON 04 JUN 2003
CA INDEXING COPYRIGHT (C) 2003 AMERICAN CHEMICAL SOCIETY (ACS)

FILE 'BIOTECHNO' ENTERED AT 12:39:46 ON 04 JUN 2003
COPYRIGHT (C) 2003 Elsevier Science B.V., Amsterdam. All rights reserved.

FILE 'CANCERLIT' ENTERED AT 12:39:46 ON 04 JUN 2003

FILE 'EMBASE' ENTERED AT 12:39:46 ON 04 JUN 2003
COPYRIGHT (C) 2003 Elsevier Science B.V.  All rights reserved.

FILE 'LIFESCI' ENTERED AT 12:39:46 ON 04 JUN 2003
COPYRIGHT (C) 2003 Cambridge Scientific Abstracts (CSA)


  10 FILES SEARCHED...
L2          27 L1


DUPLICATE IS NOT AVAILABLE IN 'FEDRIP'.
ANSWERS FROM THESE FILES WILL BE CONSIDERED UNIQUE
```